Predmet: **Mašinsko učenje 1**

Odabir i analiza baze podataka

Tema: **Procena vrednosti fudbalera sa Transfermarkt-a primenom regresionog modela**

Autori

Miloš Sirar IN 3/2020

Vasilije Zeković IN 4/2020

Datum: 15.12.2023.

Sadržaj

1 Kratak opis baze podataka

Skup podataka se sastoji od 9 .csv fajlova koji su vezani za fudbalsku analizu i vrednosti fudbalera. Sadrži podatke sa uglednog i reprezentativnog sajta Transfermarkt, odakle se podaci automatski svakodnevno ažuriraju na sajtu Kaggle. Verzija koja će se koristiti je preuzeta 1.12.2023. Poseduje informacije oko 30 hiljada fudbalera, 426 klubova, 43 lige, preko 65 hiljada utakmica i još mnogo podataka vezanih za fudbal. Svaki od .csv fajlova sadrži detaljne podatke vezane za dati entitet.

Naziv baze podataka: Football Data from Transfermarkt

Link ka bazi podataka: https://www.kaggle.com/datasets/davidcariboo/player-scores/

2 Odgovori na pitanja za analizu podataka

Ovde su dati odgovori na 16 pitanja iz prvog dela projektnog zadatka – analize baze podataka (first_milestone.pdf).

1. Problem koji će se rešavati u okviru projekta jeste procena vrednosti fudbalera u pet najjačih fudbalskih liga Evrope: Engleska, Španija, Italija, Nemačka i Francuska. Procena će se raditi na osnovu različitih obeležja vezanih za fudbalere, njihove statisitike, vrednosti klubova i ostalih parametara. Problem će se rešavati primenom više regresionih modela koji će nam dati vrednosti fudbalera kroz njihove karijere tokom perioda između 2012. i 2023. godine. Na krajuće biti izabran model koji daje optimalne procene.
2. U bazi podataka ima 42 026 uzoraka.
3. Jedan uzorak u bazi predstavlja jednu torku koja ima sve podatke za jednog fudbalera i njegov klub u trenutku procene tržišne vrednosti fudbalera.
4. U bazi ima 30 obeležja.
5. Ovde je u tabeli prikazan spisak obeležja. Dati su nazivi obeležja u bazi podataka, njihovi nazivi na srpskom i kratak opis svakog obeležja.

Attribute name	**Naziv obeležja**	**Opis obeležja**
valuation_date	Datum valuacije	*Datum procene vrednosti fudbalera*
player_id	Id fudbalera	*Identifikaciono obeležje fudbalera*
first_name	Ime	*Ime fudbalera*
last_name	Prezime	*Prezime fudbalera*
date_of_birth	Datum rođenja	*Datum rođenja fudbalera*
height_in_cm	Visina u cm	*Visina fudbalera izražena u cm*

citizienship	Nacionalnost	*Država iz koje je fudbaler*
position	Pozicija	*Deo terena na kojem fudbaler igra (golman, odbrana, vezni red, napad)*
sub_position	Konkretna pozicija	*Tačna pozicija na kojoj fudbaler igra (primer: centralni vezni)*
foot	Jača noga	*Koja noga fudbalera je jača (primarna)*
current_club_name	Naziv trenutnog kluba	*Naziv kluba za koji fudbaler trenutno igra*
current_club_domestic_competition	Domaće takmičenje trenutnog kluba	*Naziv domaćeg takmičenja trenutnog kluba*
contract_expiration_date	Datum isteka ugovora	*Datum isteka trenutnog ugovora fudbalera*
agent_name	Naziv agenta	*Naziv agencijske kuće koja zastupa fudbalera*
market_value_on_valuation_date	Tržišna vrednost na datum valuacije	*Tržišna vrednost fudbalera u trenutku valuacije*
games_played_in_domestic_competition_on_valuation_date	Broj utakmica odigranih u domaćem takmičenju na datum valuacije	*Ukupan broj utakmica koje je fudbaler odigrao u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
average_minutes_in_domestic_competition_on_valuation_date	Prosečan broj minuta u domaćem takmičenju na datum valuacije	*Prosečan broj minuta koje je fudbaler odigrao u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
goals_in_domestic_competition_on_valuation_date	Broj golova u domaćem takmičenju na datum valuacije	*Ukupan broj golova koje je fudbaler postigao u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
assists_in_domestic_competition_on_valuation_date	Broj asistencija u domaćem takmičenju na datum valuacije	*Ukupan broj asistencija koje je fudbaler ostvario u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
yellow_cards_in_domestic_competition_on_valuation_date	Broj žutih kartona u domaćem takmičenju na datum valuacije	*Broj žutih kartona koje je fudbaler prikupio u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
red_cards_in_domestic_competition_on_valuation_date	Broj crvenih kartona u domaćem takmičenju na datum valuacije	*Broj crvenih kartona koje je fudbaler prikupio u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
games_played_in_european_competition_on_valuation_date	Broj utakmica odigranih u evropskim takmičenjima na datum valuacije	*Ukupan broj utakmica koje je fudbaler odigrao u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
average_minutes_in_european_competition_on_valuation_date	Prosečan broj minuta u evropskim takmičenjima na datum valuacije	*Prosečan broj minuta koje je fudbaler odigrao u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
goals_in_european_competition_on_valuation_date	Broj golova u evropskim takmičenjima na datum valuacije	*Ukupan broj golova koje je fudbaler postigao u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
assists_in_european_competition_on_valuation_date	Broj asistencija u evropskim takmičenjima na datum valuacije	*Ukupan broj asistencija koje je fudbaler ostvario u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
yellow_cards_in_european_competition_on_valuation_date	Broj žutih kartona u evropskim takmičenjima na datum valuacije	*Broj žutih kartona koje je fudbaler prikupio u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
red_cards_in_european_competition_on_valuation_date	Broj crvenih kartona u evropskim takmičenjima na datum valuacije	*Broj crvenih kartona koje je fudbaler prikupio u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
club_name_on_valuation_date	Naziv kluba u trenutku valuacije	*Naziv kluba za koji je fudbaler igrao u trenutku valuacije*

club_domestic_competition_on_valuation_date	Domaće takmičenje kluba u trenutku valuacije	*Naziv domaćeg takmičenja kluba u trenutku valuacije*
club_value_on_valuation_year	Vrednost kluba u godini valuacije	*Vrednost kluba u godini za klub za koji je fudbaler igrao u trenutku valuacije*

6. Numeričkih obeležja ima 19.
7. Kategoričkih obeležja ima 11 u bazi podataka.
   Kategoričko obeležje koje ima najmanje različitih kategorija je foot sa tri kategorije: right, left, both.
   Kategoričko obeležje koje ima najviše različitih kategorija je last_name sa 2 180 različitih kategorija.
8. Obeležje koje se predviđa u našem modelu je: market_value_on_valuation_date. Opseg vrednosti za to obeležje je od 10 000 do 200 000 000. Prosek vrednosti je 9 624 461, dok je medijana 4 000 000.
9. Za sada ne smatramo da treba da se izbaci neko obeležje, ali to ne znači da neće biti izbačeno tokom daljeg toka projekta.
10. U bazi postoje nedostajuće vrednosti za 3 obeležja.
    Obeležje first_name – 2 023 nedostajućih vrednosti, odnosno 4.81%
    Obeležje contract_expiration_date – 1 166 nedostajućih vrednosti, odnosno 2.77%
    Obeležje agent_name – 11 408 nedostajućih vrednosti, odnosno 27.14%
11. U bazi ne postoje nevalidne vrednosti.
12. To što imamo nedostajuće vrednosti u bazi za tri obeležja, ne znači da je to greška. Na primer, mnogim brazilskim ili španskim fudbalerima su upisana samo prezimena, po kojima su prepoznatljivi i zato se ne upisuje ime. Datum isteka ugovora nije poznat za neke fudbalere, što je moguće, pa ćemo to takođe ostaviti, kao i naziv agenta.
13. Nakon svih sređivanja baze podataka, ostalo je 42 026 uzoraka.
14. Sva numerička obeležja poseduju autlajere. Autlajeri svih obeležja predstavljaju stvarne ekstremne vrednosti stoga ćemo ih zadržati.
    U pitanju su sledeća obeležja:
    1) player_id
    2) height_in_cm
    3) market_value_on_valuation_date
    4) games_played_in_domestic_competition_on_valuation_date
    5) average_minutes_in_domestic_competition_on_valuation_date
    6) goals_in_domestic_competition_on_valuation_date
    7) assists_in_domestic_competition_on_valuation_date
    8) yellow_cards_in_domestic_competition_on_valuation_date
    9) red_cards_in_domestic_competition_on_valuation_date
    10) games_played_in_european_competition_on_valuation_date
    11) average_minutes_in_european_competition_on_valuation_date
    12) goals_in_european_competition_on_valuation_date

13) assists_in_european_competition_on_valuation_date
14) yellow_cards_in_european_competition_on_valuation_date
15) red_cards_in_european_competition_on_valuation_date
16) club_value_on_valuation_year

15. Postoje dva para obeležja čiji je koeficijent korelacije veći od 0.7:
1) games_played_in_domestic_competition_on_valuation_date – average_minutes_in_domestic_competition_on_valuation_date
2) games_played_in_european_competition_on_valuation_date – average_minutes_in_ european_competition_on_valuation_date

16. Koeficijent asimetrije je 3.43, a koeficijent spljoštenosti je 18.05.

3 Napomene kako se došlo do konačne baze podataka

3.1 Napomene za svaku pojedinačnu bazu

Ovde se nalaze napomene kako je svaki od ovih 9 csv fajlova, odnosno baza podataka, iskorišten da se dobije konačna baza.

1. **appearances.csv** – 1 507 351 podataka i 13 obeležja
   Ovde su bili sadržani podaci o nastupu i učinku nekog fudbalera na jednoj utakmici. Odavde smo koristili podatke date da prepoznamo za koji valuation_date važi ova statistika, player_club_id je korišten da se pronađe za koji klub je fudbaler igrao u tom trenutku, dok su yellow_cards, red_cards, goals, assists i minutes_played korišteni da bi se našli statistički podaci za:
   1) games_played_in_domestic_competition_on_valuation_date
   2) average_minutes_in_domestic_competition_on_valuation_date
   3) goals_in_domestic_competition_on_valuation_date
   4) assists_in_domestic_competition_on_valuation_date
   5) yellow_cards_in_domestic_competition_on_valuation_date
   6) red_cards_in_domestic_competition_on_valuation_date
   7) games_played_in_european_competition_on_valuation_date
   8) average_minutes_in_european_competition_on_valuation_date
   9) goals_in_european_competition_on_valuation_date
   10) assists_in_european_competition_on_valuation_date
   11) yellow_cards_in_european_competition_on_valuation_date

12) red_cards_in_european_competition_on_valuation_date

2. **club_games.csv** – 130 432 podatka i 11 obeležja
   Ovde su sadržani podaci o odigranim mečevima između dva kluba. Ove podatke nismo koristili zato što nisu potrebni za obuku našeg modela.
3. **clubs.csv** – 426 podatka i 16 obeležja
   Ovde se nalaze detaljne informacije o samom klubu. Odavde smo koristili podatke club_name, domestic_competition_id, koje smo pomoću club_id spojili sa fudbalerima.
4. **competitions.csv** – 43 podatka i 10 obeležja
   Ovde se nalaze detaljne informacije o nazivima takmičenja i kojim državama oni pripadaju. Odavde smo koristili podatke domestic_league_code da bismo odredili da li je takmičenje domestic ili european, što je kasnije bitno za nastupe fudbalera.
5. **game_events.csv** – 666 558 podataka i 10 obeležja
   Ovde su sadržani podaci o golovima, asistencijama, na koji način su oni postignuti, u kom minutu, odnosno svaki mogući scenario na terenu. Za potrebe našeg modela, nismo hteli da ovi podaci učestvuju u obuci, jer smatramo da nisu značajni.
6. **game_lineups.csv** – 119 133 podatka i 9 obeležja
   Ovde su sadržani podaci koji govore o startnim postavama timova, odnosno ko je bio u startnoj postavi na kom meču. Ovo takođe nije bilo relevantno za potrebe obučavanja našeg modela.
7. **games.csv** – 65 216 podataka i 23 obeležja
   Ovde su sadržani podaci slično kao i kod club_games, samo sa nekim malo drugačijim obeležjima, ali takođe je bilo nerelevantno za obučavanje modela koji mi želimo da postignemo.
8. **player_valuations.csv** – 440 663 podataka i 9 obeležja
   Ovde su sadržani podaci o tome koliko su fudbaleri vredili u određenom vremenskom periodu. Obeležje date je korišteno kao valuation_date, market_value_in_eur je korišten kao market_value_on_valuation_date, dok je player_id korišten da bi se moglo povezati sa potrebim obeležjima iz ostalih tabela. Ostala obeležja iz ove tabele su nepotrebna za ovu analizu.
9. **players.csv** – 30 302 podatka i 23 obeležja
   Ovde su sadržani podaci o osnovnim odlikama fudbalera. Player_id, first_name, last_name, date_of_birth, position, sub_position, foot, height_in_cm, contract_expiration_date, agent_name su obeležja koja smo koristili u našoj tabeli. Takođe obeležja country_of_citizenship smo preimenovali u citizenship. Obeležja current_club_id smo iskoristili da bismo pronašli current_club_name, current_club_domestic_competition_id smo iskoristili da bismo pronašli current_club_domestic_competition. Ostala obeležja nam nisu bila potrebna.

3.2 Napomene prilikom procesa spajanja baza

U ovom poglavlju je navedeno još par bitnih napomena za proces spajanja baza. Ovde je navedeno nekoliko najvažnijih koraka koji su urađeni da bi se došlo do trenutne baze podataka.

1. Vrednosti fudbalera se procenjuju od 09.07.2012. godine do 01.12.2023. zato što nastupi fudbalera postoje samo od 09.07.2012., a 01.12.2023. smo preuzeli podatke.
2. Iz tabele players su izbačeni svi fudbaleri za koje nije postojao podatak foot ili height_in_cm, zato što mislimo da su to dosta bitni podaci za naš model, a mislimo takođe da je neprikladno da fudbalerima dodeljujemo visinu na osnovu proseka ili pomoću neke druge tehnike, jer nije relevantno. Takođe neki fudbaleri igraju levom nogom, neki desnom, a neki podjednako dobro sa obe noge, pa je takođe loše da tu stavljamo neku srednju vrednost, to može značajno da utiče na cenu fudbalera.
3. Sve ovo je spajano u više koraka da bi se na kraju dobila ova konačna tabela.
4. Razne statistike fudbalera u periodu valuacije koje su gore navedene su pronađene tako što su se tabele appearances, players i player_valuations spajale kroz cikluse. Gledala se statistika fudbalera između dva perioda procena vrednosti fudbalera za te datume valuacija, i dodavala se u ukupnu statistiku za taj period. Isto ovo je urađeno za svaki period valuacije koji postoji tokom vremena za kada se procenjuju vrednosti fudbalera.
5. Ukoliko fudbaleru nije bio poznat klub, zbog toga što nije imao nastupe u toku jednog perioda valuacije, uzimali smo klub iz naredne valuacije. Ukoliko ni to nije bilo poznato, fudbaler je tada bio izbačen iz skupa podataka, jer ne znamo za koji klub je igrao i samim tim nije relevantan za našu procenu.
6. club_value_on_valuation_year – ovo obeležje je dobijeno nakon prolaska kroz više tabela i pomoćnih tabela, tako što smo gledali koji fudbaleri su imali valuacije u toj godini i igrali za taj klub. Ukoliko je više valuacija bilo u toku te godine za tog fudbalera, uzimali smo prosek tržišne vrednosti fudbalera. Zatim smo sabrali vrednosti svih fudbalera koji su igrali u jednom klubu u jednoj godini i to predstavlja vrednost kluba u toj godini.